DOCUMENT RESUME

ED 307 291 TM 013 284

AUTHOR Anderson, Ronald W.
TITLE The Effects of Group-Based Mastery Learning and Enhanced Cognitive Entry Behaviors on Algebra Achievement.
PUB DATE Mar 89
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Algebra; Effect Size; Grade 9; *Group Instruction; High School Freshmen; Instructional Effectiveness; *Mastery Learning; *Mathematics Achievement; Mathematics Tests; *Norm Referenced Tests; Secondary Education; Secondary School Students; Standardized Tests; *Teacher Made Tests
IDENTIFIERS *Cognitive Entry Behaviors

ABSTRACT

This quasi-experiment was designed to assess Robert Slavin's challenge to research on the effects of mastery learning on student achievement in algebra. Focus was on determining if there would be a difference in mathematics achievement between an experimental group (EG) of students whose initial cognitive entry skills were enhanced and who were subsequently taught under mastery learning conditions and a control group (CG) of similar students who were taught by conventional instruction. Applying stringent criteria for internal and external validity, Slavin's literature analysis using the technique of best evidence synthesis found only moderately positive effects of mastery learning on experimenter-made tests and virtually no evidence to support the effectiveness of mastery learning on standardized tests. In the present study, there were two EG classes (morning and afternoon) of ninth grade algebra I students (n=46) and two CG classes (morning and afternoon) of ninth graders (n=40). The subject groups were equivalent in mathematics achievement at the beginning of the experiment. All groups were administered the Orleans-Hanna Algebra Prognosis Test during the first few days of the experiment. The prerequisite mathematics skills of the EG students were remediated during the first week of the experiment. The EGs were taught for 18 weeks under mastery learning conditions. The CGs received traditional instruction for 18 weeks. During the final week of the experiment, all groups were given a teacher-made test and a standardized, norm-referenced test--the Step III Algebra End-of-Course Test. Teacher-made test results support Slavin's conclusions, while standardized test results indicate differences that exceed the range of effect sizes reported by Slavin. Results also show that more time was required in the morning class to remediate deficiencies in prescribed skills, and more time was available for enrichment activities in the afternoon classes. (TJH)

ED307291

The Effects of Group-Based Mastery Learning and Enhanced Cognitive Entry Behaviors on Algebra Achievement

THE EFFECTS OF GROUP-BASED MASTERY LEARNING AND ENHANCED COGNITIVE ENTRY BEHAVIORS ON ALGEBRA ACHIEVEMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald W. Anderson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Ronald W. Anderson, Ed.D.
Greene County Office of Education
360 East Enon Road
Yellow Springs, Ohio 45387

A paper presented to the American Educational Research Association, San Francisco, California, March 30, 1989.

TM013284

BEST COPY AVAILABLE

ABSTRACT

For over a decade, only compellingly positive effects of mastery learning on student achievement were reported in the literature. Recently however, Robert Slavin has challenged the analytical techniques used to make those claims. Applying stringent criteria for internal and external validity, Slavin's literature analysis using the technique of best evidence synthesis found only moderately positive effects of mastery learning on experimenter-made tests and virtually no evidence to support the effectiveness of mastery learning on standardized tests.

In looking at the effect of mastery learning on algebra achievement, the design of this study incorporated what Slavin proposed to be the "best evidence" for mastery learning. The quasi-experiment included two experimental groups and two control groups, taught by different teachers. The groups were equivalent in mathematics achievement at the beginning of the experiment. All groups were given the Orleans-Hanna Algebra Prognosis test during the first few days of the experiment.

The prerequisite mathematics skills of the experimental group students were remediated during the first week of the experiment. The experimental groups were taught for eighteen weeks under mastery learning conditions. The control groups received traditional instruction. During the final week of the experiment, all groups were given both a

teacher-made test as well as a standardized, normative-referenced test: the Step III Algebra End-of-Course Test.

Both of the experimental groups performed better on the teacher-made algebra test than their control group counterparts, with effect sizes of .274 and .856. On the standardized posttest, the afternoon experimental group out-performed its control group yielding a relatively high effect size of .724. However, the morning control group performed better than its counterpart, with an effect size of -.181.

The effect sizes found in this experiment on the teacher-made posttest (.274 and .856) parallel closely the range of effect sizes reported by the Slavin best-evidence synthesis: .18 to .90. The effect sizes found in the experiment on the standardized posttest (-.181 and .724) exceed the range of effect sizes reported by Slavin: zero to .25.

The Effects of Group-Based Mastery Learning and Enhanced Cognitive Entry Behaviors on Algebra Achievement

INTRODUCTION

The long-established, time-tested effectiveness of group-based mastery learning as an instructional strategy to enhance student achievement has been challenged. Until recently, only compellingly positive effects of mastery learning on achievement were reported in the literature.

As recently as three years ago, meta-analyses of the research by Guskey and Gates (1985, 1986) as well as by Kulik, Kulik, and Bangert-Drowns (1986) touted the effects of group-based mastery learning. Moreover, just over a decade ago, a full scale review of the literature by Block and Burns (1976) also claimed extraordinarily positive effects for group-based mastery learning. Seemingly, the success of mastery learning had stood the test of time. That is, until now.

In striking contrast to earlier literature reviews, Robert Slavin (1987) has reviewed again the mastery learning literature through the lenses of his technique of best evidence synthesis (Slavin, 1986). As a result of this synthesis of the literature, the efficacy of mastery learning is now in question.

Slavin's best-evidence synthesis reflects only moderately positive effects of mastery learning on experimenter made measures and virtually no evidence to support the effectiveness of mastery learning on standardized measures of achievement.

Accordingly, Slavin's synthesis of the literature brings to light a significant educational dilemma in the polarity of the "coverage" versus "mastery" issue. Slavin claims that students who are taught under mastery learning conditions become competent in a narrowly defined, selective set of learning objectives (e.g. mastery) at the expense of a proficiency in the breadth (e.g. coverage) of the material that is intended to be taught. In other words, students who are taught under mastery learning conditions will perform better on teacher-made (criterion-referenced) measures of achievement than students who are taught under conventional conditions. On the other hand, students taught under conventional learning conditions will perform better on standardized (normative-referenced) measures of achievement than students taught under mastery learning conditions.

This dilemma is extraordinarily significant in light of the impetus by many states to mandate both the reporting of criterion-referenced competency test scores as well as standardized test results as indicators of public accountability.

Background

The theory of mastery learning came about as the result of efforts by educators to find instructional methods that would enable virtually all children to learn the content presented to them. This effort was begun as early as the 16th century in European Jusuit schools (Kuhn, 1985). Later, the ideas of other early educators, including Comenius, Pestalozzi, and Herbart also helped lay the foundation for later work in the area of mastery learning (Bloom, 1974). In recent history, educational researchers have been interested in factors that influence

student learning in and outside of the classroom. Much of the recent research effort has been focused not upon the characteristics of good teachers, but rather upon the instructional behaviors displayed by classroom teachers that result in higher levels of student achievement. (Bloom, 1980). As a result, a great deal has been learned about not only specific teacher behaviors that enhance learning, but also about instructional conditions that foster improved student achievement.

John B. Carroll's (1963) Model for School Learning provided the modern-day impetus for renewal in the belief that all children can learn what is taught to them. Carroll redefined the concept of aptitude to be not so much an innate indicator of student success as a function of time needed by the student to learn a particular concept or skill.

In the 1960's, Benjamin Bloom (1968) extended Carroll's perspective on aptitude defined as a function of time needed to learn. The extension of these ideas became his theory of learning for mastery. Mastery learning theory takes into account a learner's cognitive entry behaviors as well as his/her affective entry characteristics. The mastery learning model suggests that when cognitive entry behaviors and affective entry characteristics are optimal, and when instruction is appropriate, most students can learn when given sufficient time to do so. In the past decade, research in the field of mastery learning has focused on the enhancement of cognitive entry behaviors and upon instruction, which is defined to include primarily the components of feedback and correctives. Group-based mastery learning has generally been found to produce student achievement at a level one full standard deviation above the mean of students taught under non-mastery con-

ditions (Walberg, 1984).

The 2 Sigma Problem

A more ambitious goal of mastery learning, though, has been to identify group-based instructional techniques and conditions that are as effective as the most effective teaching technique presently known -- one-to-one tutoring. It has been shown that students taught by one-to-one tutoring have achieved at a level two standard deviations above the average student taught by conventional methods (Walberg, 1984). Having shown that students taught under mastery learning conditions can perform one standard deviation above the mean of the average students taught under conventional conditions, Benjamin Bloom (1984a, 1984b) and his students began the task of finding instructional strategies that, when used in tandem with mastery learning, could approach the level of achievement currently attainable by one-to-one tutoring. The problem became known as the 2 sigma problem.

Best Evidence Synthesis Challenges Those Claims

When Slavin examines what he considers to be the best evidence for mastery learning, the achievement claims of the theory's proponents are called into question. Slavin (1987) states:

> ...the 2 sigma challenge proposed by Bloom (1984a, 1984b) is probably unrealistic, certainly within the context of group-based mastery learning. Bloom's claim that mastery learning can improve achievement by more than 1 sigma is based on brief, small, artificial studies that provided additional instructional time to the experimental classes. ...The 2 sigma challenge (or 1-sigma claim) is misleading out of context and potentially damaging to educational research both within and outside of the mastery learning tradition, as it may lead researchers to belittle true, repli-

> cable, and generalizable achievement effects in the more realistic range of 20-50% of an individual-level standard deviation.

Slavin criticizes the inclusion of certain previous mastery learning studies in literature reviews and meta-analyses on the grounds that:

1. the students in experimental mastery learning classes received more instructional time than did students in the control group;
2. tests used to measure achievement were constructed by the experimenter and therefore biased toward the experimental group; and
3. mastery learning studies were of short duration (less than four weeks).

Therefore, it became readily apparent that further research could be of value to test the efficacy of mastery learning under conditions described by Slavin that would provide the "best evidence possible." The purpose of this study, then, was to address both the effectiveness of group-based mastery learning, which is in question, and to explore the coverage verses mastery dilemma under "best evidence conditions."

Statement of the Research Objective

Based upon evidence presented in the literature, it was anticipated that the theory of learning for mastery would be supported when tested experimentally. The objective of this research was to determine if there would be a difference in mathematics achievement, as measured by both a criterion-referenced and a normative-referenced test, between an experimental group of students whose initial cognitive entry skills were enhanced and who were subsequently taught under mastery learning conditions and a control group of similar students who were taught by conventional instruction.

The research objective was clarified using the statistic of effect size. Effect size is the z-score of the mean of the experimental group referenced in the frequency distribution of the control group, that is the mean of the experimental group minus the mean of the control group expressed in standard deviation units (Glass, 1981). Burns (1979) reports that, in general, an effect size of .20 indicates a weak effect, an effect size of .50 indicates a moderate effect, and an effect size of .80 indicates a strong effect.

Overview of the Study

The research question was tested using a quasi-experimental, non-equivalent control group design. There were two experimental group classes of ninth grade Algebra I students (one in the morning and one in the afternoon) and two control group classes (again, a morning class and and afternoon class). The experimental groups and the control groups were taught by different teachers and were in different school buildings. The experimental groups and the control groups were selected from an accessible sample of students.

The experimental group received a diagnostic pretest (Orleans-Hanna Algebra Prognosis Test) on the second day of school in the fall. Then students' cognitive entry behaviors were enhanced. Based on the results of the Orleans-Hanna, students' prerequisite Algebra skill weaknesses were identified and deficits were remediated during the first week of school. This was followed by 18 weeks of instruction in Algebra I under mastery learning conditions. The experimental group received both a teacher-made criterion-referenced posttest and a normative-referenced posttest (Step III Algebra Test).

The control groups, on the other hand, received the pretest and both posttests, but cognitive entry behaviors were not enhanced during the first week of school. The control groups were taught for 18 weeks by conventional group instruction.

The dependent variable in the study was achievement in Algebra I as measured by both the teacher-made criterion-referenced test and by the nomrative-referenced standardized achievement test. The independent variables were the treatments: the enhancement of students' cognitive entry behaviors and instruction under mastery learning conditions.

The design of the quasi-experiment incorporated all nine "conditions" established by Slavin to represent the best evidence for mastery learning.

1. Best evidence conditions require equal instructional time and equal objectives. In this study, instructional time for the experimental and control groups was held constant. Both teachers followed the objectives prescribed in their local mathematics course of study and used the same textbook.

2. Best evidence conditions require students to be tested on their mastery of instructional objectives at least once every four weeks. The students in this study's experimental groups were tested eight times during the 18 weeks of the experiment, or once every 2.25 weeks.

3. Best evidence conditions demand students to be taught as a total group prior to formative testing. Such was the design of this study.

4. Best evidence conditions permit mastery learning to be the only or the primary intervention. The students in this study's experimental groups participated in activities designed to enhance their cognitive entry behaviors (an integral part of mastery learning theory) and were taught under mastery learning conditions.

5. Best evidence conditions reflect group-based mastery learning as compared to traditional instruction not using the feedback-corrective cycle. This study met that requirement.

6. Best evidence conditions require evidence that the experimental and control groups were initially equal. In this study, the Orleans-Hanna pretest was used to determine this equality.

7. Best evidence conditions allow only studies of four weeks or longer. This study was conducted for 18 weeks.

8. Best evidence conditions require that at least two experimental and two control classes and/or teachers are involved. This study included two experimental and two control classes taught by separate teachers.

9. Finally, best evidence conditions require that the achievement measures reflect the content taught by both the experimental and control groups. In this study, the items included on the teacher-made posttest were contributed by the teachers of both groups and reflected the content taught by both teachers.

Limitations of the Study

Care should be exercised in generalizing the results of this study in that the experiment was conducted in only one content area (Algebra I) at only one grade level (9). Moreover, the sample was drawn from an accessible population of ninth graders and both the teacher of the experimental groups and the teacher of the control groups were volunteers. Although not specifically a limitation, it should be noted that 24.56% of the subjects in the experiment, for one reason or another, could not be included in the reporting of final results. This rather high mortality rate reflects students who withdrew from the school system, students who were moved to other classes, and students who were unavailable for any of the pre- or posttests. Only students who were available for all of the tests were included in the data analysis. The limitation of the relatively high mortality rate, however, is somewhat neutralized by the benefits of conducting the experiment in a "real world" setting.

The table below lists the number of students for whom data is reported and illustrates the percentage of mortality in each group:

Experimental Mortality

Research Group	n: Beginning	n: End	Mortality
Experimental [A.M.]	32	21	34.38%
Experimental [P.M.]	29	25	13.79%
Total Experimental	61	46	24.59%
Control [A.M.]	26	19	26.92%
Control [P.M.]	27	21	22.22%
Total Control	53	40	24.53%
GRAND TOTALS	114	86	24.56%

As the table indicates, there were 114 students in the experiment at the beginning of the school year. However, due to student movement in and out of classes, tests scores on all three tests were available for 86 students. This reflects an attrition rate of 24.56 percent.

Again, only students taking the pretest and both posttests were included in the analysis of the data.

Description of the Enhancement of Cognitive Entry Behaviors

The first experimental treatment consisted of enhancing the cognitive entry behaviors of the experimental group. The enhancement of cognitive entry behaviors is an instructional intervention designed to remediate any deficiencies in prerequisite mathematics skills before

instruction is begun. Following the administration of the Orleans-Hanna Algebra Prognosis Test, an item analysis was prepared for the experimental group teacher by the experimenter. The analysis reported the number of students missing each item, as well as the names of the students. This analysis facilitated grouping of students for remediation.

The third and fourth days of school were spent remediating mathematics skills which were identified for each student by the pretest. The students in the experimental class were given feedback detailing the number of and types of items which were answered incorrectly on the pretest. Corrective instruction consisted of both individualized and group-based remedial instruction. Several instructional strategies were employed. For individual students, paper and pencil worksheets provided further drill and practice. The worksheets were correlated by the teacher to the students' skill deficiencies. Whole group instruction provided additional practice of the deficient skills and was carried out with individual students or small groups of students solving problems at the chalkboard.

Using these strategies, the instructor was able to point out the students' procedural or logical errors, provide the students with an opportunity to re-try the problems under the instructor's supervision, and provide an opportunity for the students to receive acknowledgment when the correct answers were obtained.

Instruction Under Mastery Learning Conditions

The second experimental treatment consisted of instruction under mastery learning conditions. This treatment was carried out for the duration of the experiment. The teacher of the experimental group followed the Block and Anderson (1975) mastery learning implementation

strategy. The experimental class was taught throughout the experiment by whole-group instruction. The content to be learned by the students was defined by the objectives prescribed in the local district mathematics course of study. The level of mastery at which each student was expected to master the objectives was set mutually by the teacher and the experimenter at the 80% level.

Algebra skills were taught to the students in the experimental group by direct instruction using the textbook as well as teacher-prepared materials. The textbook used in the experimental class was the Houghton-Mifflin Algebra: Structure and Method, Book I. The experimental group teacher first presented a new skill or concept to the students. Students were shown how to perform the skill and were told how the skill related to what was learned previously. Then, students were given an opportunity to practice the new skill under the guidance of the instructor. This was accomplished by having groups of students go to the chalkboard or by having students try the skill at their seats. During this segment of instruction, students were provided with immediate feedback concerning errors and with corrective instruction, if necessary.

Students were then given a home-work assignment which consisted of further practice in the algebra skill. The home-work assignments were gone over at the beginning of the next class period and additional feedback and correctives were provided as necessary.

Following short units of instruction, usually chapter units defined by the textbook, students were given a formative test correlating to the prescribed objectives for the unit. The formative test was corrected by the teacher, but the students' scores did not influence their

final grade in the class. Each student's formative test provided feedback to him or her concerning any remaining skill deficiencies.

Following the administration of the formative test, students who did not achieve at the 80% level participated in remedial and corrective activities correlated by the teacher to individual skill deficiencies. Students who scored at the 80% level or better were given enrichment work. The enrichment activities consisted of further application of the skill using more complex (harder-to-solve) problems.

Following one cycle of the feedback-corrective process, students were administered a summative posttest which did count toward the students' final grades in the class. The summative test consisted of a second form of the formative test, thus testing the same skills and concepts. Both the formative and summative tests were teacher made. However, items were frequently drawn from materials provided by the textbook publisher.

Instruction Under Conventional Conditions

Conventional instruction for the control group consisted of traditional forms of mathematics instruction. The instructional strategies employed by the teacher of the control groups were similar to those used by the experimental group's teacher. However, students in the control group did not participate in activities designed to enhance cognitive entry behaviors, nor did students participate in cycles of formative testing, correctives, feedback, and summative testing.

The content to be learned by the students in the control groups was, like the experimental group, defined by the local district's mathematics course of study. The control group students also used the Houghton-Mifflin Algebra: Structure and Method, Book I.

As with the experimental group, the teacher of the control group classes introduced new algebra skills and concepts through direct instruction. After presentation of the new skill or concept, students participated in guided practice either at the chalkboard or at their seats. Further drill and practice was provided through home-work assignments. Home-work assignments were corrected in class and students were given an opportunity to ask questions about items they missed on the home-work assignment.

Although the control group teacher was not using mastery learning strategies, students were given feedback and corrective instruction during the period of in-class guided practice and again following correction of the home-work assignment. The feedback and correctives given during guided practice and following correction of the home-work assignment were very similar for both the experimental and control groups. The following differences between instruction in the experimental and control groups should be noted:

1. Students in the control group were not provided with an opportunity to remediate prerequisite mathematics skills prior to instruction in Algebra I. For this group, any skill deficiencies that were carried over from eighth grade mathematics remained uncorrected as instruction in Algebra began.

2. Feedback and correctives were provided to the experimental group in a formal and structured manner. The results of the formative test enabled the instructor to identify each student who did not perform at the 80% level and corrective instruction was targeted toward identified skill weaknesses. In the control groups, feedback and correctives were provided only to students who were randomly sent to the chalkboard during guided practice or to those students who volunteered questions following the correction of the home-work assignment.

Time Allocation and Content Covered

Instructional time for both the experimental and control groups was limited to the amount of time allocated during the school day in the district for ninth grade Algebra I. The amount of time taken for formative testing and corrective instruction for the experimental group was taken from regularly scheduled class time. No additional time was provided for out-of-class remediation. As a result, more instructional time was available to the control group within the period of the experiment. The duration of the experiment was 18 weeks.

Although both the experimental and control groups used the same course of study and the same textbook, the control group was able to cover more content than the experimental group.

Description of Tests Used

The Orleans-Hanna Algebra Prognosis Test consists of 60 multiple-choice items and five self-report items. The authors report the predictive validity of the test scores to correlate .73 with mid-year algebra grades and .83 with end-of-the-year algebra achievement test scores. A correlation of .72 is reported between the Orleans-Hanna test and end-of-the-year algebra grades. The authors report a reliability coefficient for the 60 multiple choice items of .96 based upon the internal consistency of an eight-grade sample. The authors also report the test's reliability in terms of the relationship between scores on the 60-item multiple choice test and the scores on the individual items. A median point-biserial correlation of .53 was reported.

The teacher-made criterion-referenced test consisted of 24 items which were contributed by both the experimental and control group teachers. The 24 items reflected the content prescribed by the objec-

tives in the local mathematics course of study. As a teacher-made test, no assessment of the validity or reliability was made.

The Step III Algebra end-of-Course Test was designed to measure student's final status achievement in algebra. The test consists of 30 multiple choice items and is published by CTB McGraw-Hill. Form X of the test was given. The test manual reports a reliability for the test of .75 as well as a point biserial correlations of .47 for item discrimination.

Findings

At the completion of the experiment, five sets of data were available which could be analyzed and from which conclusions about the experiment could be drawn: 1. the Orleans-Hanna Algebra Prognosis Test; 2. eight formative tests; 3. eight summative tests; 4. the teacher-made criterion-referenced algebra test; and 5. the Step III Algebra End-of-Course Test.

The data sets were available for four groups of students: the two experimental groups and the two control groups. The groups are designated one through four as follows:

Group 1: A.M. Experimental Group	21 Subjects
Group 2: A.M. Control Group	19 Subjects
Group 3: P.M. Experimental Group	25 Subjects
Group 4: P.M. Control Group	21 Subjects

In order to analyze the pre- and posttest data, two statistical procedures were employed. First, tests of statistical significance were performed on the results of the Orleans-Hanna Algebra Prognosis Test to determine if the experimental and control groups were equal going into the experiment. Secondly, in order to compare the end-status achieve-

ment of the experimental and control groups, effect sizes were computed on the criterion-referenced algebra test and on the Step III Algebra End-of-Course Test.

Analysis to Determine Pre-Test Differences

Using the raw score data from the Orleans-Hanna Algebra Prognosis Test, a t-test was performed first on groups 1 and 2 and then on groups 3 and 4 to determine if there was a significant difference in the means of the two groups.

The data in Table 1 below show the results of the t tests for the Orleans-Hanna pretest. The t-value for the morning experimental and control groups (groups 1 and 2) was 1.14. This represents a 2-tailed probability, or observed significance level, of .263. This is larger than the .05 level of significance, and therefore, the hypothesis that the two population means are equal must be accepted. We conclude, therefore, that the A.M. experimental and control groups were equivalent at the onset of the experiment.

Table 1.
Tests of Statistical Significance for Orleans-Hanna PreTest Means

	Experimental			Control			
Group	n	M	SD	n	M	SD	t
A.M.	21	90.9048	4.929	19	88.0800	5.908	1.14 ns
P.M.	25	89.2632	4.121	21	86.5714	1.182	.92 ns

The t-test for the afternoon experimental and control groups (Groups 3 and 4) yielded a t-value of .92. This represents a 2-tailed probability of .360. Since this observed significance level is also larger than the .05 level of significance, we can, as with the morning groups, conclude that the two population means are equal. Therefore, we can say also that both the afternoon experimental and control groups were equivalent in algebra achievement at the beginning of the experiment.

Analysis of Data to Address the Research Objective

The results of the teacher-made criterion-referenced algebra test and the standardized normative-referenced Step III Algebra End-of-Course Test were analyzed in order to answer the research objective. The statistic of effect size was computed for each.

Method Used to Compute Effect Size

Since both *t* tests performed on the Orleans-Hanna pre-test proved to be nonsignificant, effect size was computed using the method suggested by Glass et al. (1981) for use when there are no differences in experimental and control group pretest means. Under these circumstances, effect size is computed by determining the difference between the posttest means of the control group and the experimental group, and then by dividing the difference by the standard deviation of the control group.

$$\text{E.S.} = \frac{\text{Experimental Group Mean Minus Control Group Mean}}{\text{Standard Deviation of the Control Group}}$$

The resulting effect size indicates to what degree the experimental group performed in relation to the control group. A positive effect size indicates that the experimental group achieved at a higher level than the control group. Conversely, a negative effect size indicates that the control group performed at a higher level than the experimental group. Thus, effect size is reported in standard deviation units, using the standard deviation of the control group as the metric. An effect size of positive .5, for example, would indicate that the average student in the experimental group scored one-half standard deviations higher than the average student in the control group. In carrying out the analysis for each of the posttests, means, standard deviations, and effect sizes were computed first for the morning experimental and control groups (Groups 1 and 2), and then for the afternoon experimental and control groups (Groups 3 and 4). Tables 2 and 3 below show the results for the two posttests.

Table 2.
Effect Sizes for the Teacher-Made Algebra Test

Group	n	M	SD	Effect Size
A.M. 1 Exp	21	19.095	1.841	
A.M. 2 Control	19	18.421	2.457	+.274
P.M. 3 Exp	25	20.040	1.859	
P.M. 4 Control	21	17.952	2.439	+.856

Table 3.
Effect Sizes for the Step III Algebra End-of-Course Test

Group	n	M	SD	Effect Size
A.M. 1 Exp	21	18.429	2.767	
A.M. 2 Control	19	18.947	2.857	-.181
A.M. 3 Exp	25	18.760	2.712	
A.M. 4 Control	21	16.619	2.958	+.724

Analysis of the Teacher-Made Algebra Test

Comparison of Morning Experimental and Control Groups

The average score of the 21 students in the morning experimental group was 19.095 on the teacher-made algebra criterion-referenced achievement measure. The standard deviation of the group was 1.841. By contrast, the average score of the 19 students in the morning control group was 18.421 and their standard deviation was 2.457. It appears that the students in the morning experimental group achieved only slightly higher level than the students in the morning control group with .674 raw score points difference between the group means.

A comparison of the means and standard deviations of the two groups yields an effect size of .274. In other words, the average student in the morning experimental group performed higher than 60% of the students in the morning control group. A comparison of the standard deviations indicates that the students in the experimental group were more homogeneous at the end of the experiment than their control group counterparts.

The contrasting homogeneities of the two morning groups on the teacher-made algebra test are illustrated by the stem and leaf display in Table 4 below. The graph shows the raw score frequencies for the experimental and control groups on the teacher-made algebra test. The distribution for the experimental group is negatively skewed. Most scores are clustered closely about the mean; two of the scores which stand far apart from the mean are lower achievers. By contrast, the distribution of the scores of the students in the morning control group is relatively symmetrical, but visually platykurtic (flat). The scores are dispersed throughout the range of scores rather than clustered about the mean. As the display shows, an equal number of students in both morning groups scored at the high end of the range (at or above 21). However, more than four times as many students in the control group scored at the low end of the range (at or below 17).

Comparison of Afternoon Experimental and Control Groups

The average score of the 25 students in the afternoon experimental group was 20.04 on the teacher-made algebra test. The standard deviation of the group was 1.859. The 21 students in the afternoon control group scored lower, with a mean of 17.952. The effect size of .856 indicates that the average student in the experimental group achieved higher than 80% of the students in the afternoon control group.

The stem and leaf display contrasts the homogeneity of the afternoon experimental and control groups. Like the morning experimental group, the distribution of the afternoon experimental group is negatively skewed with most scores clustered closely about the mean.

Table 4.
Stem and Leaf Display of Raw Score Frequencies on the 24-item Teacher-Made Algebra Posttest

Unit = 1.0	
1	
1	4
1	
1	6
1	
1	8 8 8 8 8
1	9 9 9 9 9
2	0 0 0 0
2	1 1 1 1
2	2
2	

A.M. Experimental

$\underline{n}$ = 21
$\overline{M}$ = 19.095
SD = 1.841
Range = 14 to 22

Unit = 1.0	
1	
1	
1	5 5
1	6 6 6
1	7 7 7 7
1	8
1	9 9 9
2	
2	1 1 1
2	2 2 2
2	

A.M. Control

$\underline{n}$ = 19
$\overline{M}$ = 18.421
SD = 2.457
Range = 15 to 22

Unit = 1.0	
1	
1	
1	
1	6
1	7 7
1	8 8 8
1	9 9
2	0 0 0 0 0
2	1 1 1 1 1 1
2	2 2 2 2 2
2	3

P.M. Experimental

$\underline{n}$ = 25
$\overline{M}$ = 20.040
SD = 1.859
Range = 16 to 23

Unit = 1.0	
1	3
1	4
1	5 5
1	
1	7 7 7 7 7
1	8 8 8 8 8
1	
2	0 0 0
2	1 1 1
2	2
2	

P.M. Control

$\underline{n}$ = 21
$\overline{M}$ = 17.952
SD = 2.439
Range = 13 to 22

The distribution of scores of the afternoon control group is also negatively skewed, but the curve is visually more mesokurtic (like a normal curve). Three times as many afternoon experimental group students fall at the high end of the spectrum of scores (at or above 21), while three times as many control group students fall at the lower end of the range (at or below 17).

Analysis of the Standardized Algebra Posttest

Comparison of Morning Experimental and Control Groups

The achievement of the students on the standardized Step III Algebra End-of-Course Test. As the data in Table 3 shows, the average score of students in the morning experimental group was 18.429. The standard deviation of the group was 2.767. By comparison, the students in the morning control group performed at a slightly higher level than the students in the morning experimental group. The mean score of the morning control group was 18.947, or .518 raw score points higher than the experimental group. The standard deviation of the morning control group (2.857) indicates that they were less homogeneous than the students in the morning experimental group.

The resulting effect size of -.181 indicates that the average student in the control group achieved higher than 57% of the students in the experimental group.

The variabilbity of the morning groups on the standardized test is shown by the stem and leaf display in Table 5. Both distributions are visually symmetrical. The means and standard deviations of the two groups are very similar with scores dispersed throughout the range. Two students in the experimental group fall at the high end of the

Table 5.
Stem and Leaf Display of Raw Score Frequencies on the Standardized Step III Algebra End-of-Course Test

Unit = 1.0

```
1
1    3
1    4
1    5 5
1    6 6
1
1    8 8 8 8 8
1    9
2    0 0 0
2    1 1 1 1
2    2
2    3
2
2
2
```

A.M. Experimental
n = 21
M = 18.429
SD = 2.767
Range = 13 to 23

Unit = 1.0

```
1
1
1    4
1    5
1
1    7 7 7 7
1    8 8 8 8
1    9 9 9
2
2    1 1 1
2    2
2    3
2
2
2    6
```

A.M. Control
n = 19
M = 18.947
SD = 2.857
Range = 14 to 26

Unit = 1.0

```
1
1
1    4 4
1
1    6 6 6
1    7 7 7 7
1    8 8 8 8
1    9 9
2    0 0 0 0
2
2    2 2 2 2
2    3
2    4
2
2
```

P.M. Experimental
n = 25
M = 18.760
SD = 2.712
Range = 14 to 24

Unit = 1.0

```
1    2
1    3 3
1    4 4 4 4
1    5 5
1
1    7 7 7 7 7
1    8
1    9 9
2    0 0
2
2    2 2
2
2
2
2
```

P.M. Control
n = 21
M = 16.619
SD = 2.958
Range = 12 to 22

distribution of scores (22 or above) while three students in the control group fall into that range. However, the groups are not as similar at the other end of the distribution; three times as many experimental group students fall at the low end of the range (16 or below).

Comparison of Afternoon Experimental and Control Groups

The mean score of the students in the afternoon experimental group was 18.760 on the standardized test. The standard deviation of this group was 2.712. On the other hand, the students in the afternoon control group scored lower, with an average score of 16.619, or 2.141 raw score points difference. They, too, were slightly less homogeneous than the experimental group with a standard deviation of 2.958.

The effect size of .724 indicates that the average student in the afternoon experimental group scored higher than 76% of the students in the afternoon control group on the Step III Algebra End-of-Course Test.

The mean of the afternoon experimental group is higher than the mean of the afternoon control group, but the distribution of scores in the control group is negatively skewed. There are three times as many experimental group students at the high end of the range (22 or above) while almost twice as many control group students at the lower end of the spectrum of scores (16 and below).

Mastery Levels of Experimental Groups on Formative and Summative Tests

Analysis of the Morning Experimental Group

Table 6 below shows the percentage of students in each of the experimental classes who achieved mastery at the 80% level on the eight for-

mative and summative tests. Also shown are the percentage of students who had not attained 80% mastery of the skills at the conclusion of the cycle. As the table shows, an average of 67% of the students in the morning experimental group achieved mastery on the formative tests, and therefore were not in need of corrective instruction. These students were provided with additional enrichment instruction. However, 33% of the students in the class, on the average, did participate in remedial

Table 6.
Percentage of Experimental Group Students who Achieved Mastery on the Formative and Summative Tests

Test	Percentage Who Achieved 80% Mastery on Formative Test	Percentage Who Achieved 80% Mastery on Summative Test	Percentage Who Did Not Achieve Mastery At 80%
	Experimental Group 1 (Morning Group)		
1	70%	92%	8%
2	92%	100%	0%
3	67%	96%	4%
4	83%	92%	8%
5	75%	92%	8%
6	63%	79%	21%
7	88%	92%	8%
8	100%	100%	0%
MEAN	67%	93%	7%
	Experimental Group 3 (Afternoon Group)		
1	46%	86%	14%
2	97%	97%	3%
3	68%	100%	0%
4	93%	97%	3%
5	86%	93%	7%
6	96%	96%	4%
7	93%	93%	7%
8	96%	96%	4%
MEAN	84%	95%	5%

and corrective activities. As a result, following the administration of the summative tests, an average of 93% of the students had attained mastery at the 80% level. Only 7% of the students, on the average, did not attain mastery before moving on to the next unit.

Analysis of the Afternoon Experimental Group

As shown in the table, an average of 84% percent of the students in the afternoon experimental class attained mastery at the 80% level on the eight formative tests. Sixteen percent of the students, on the average, were in need of further remedial instruction. Following the administration of the summative tests, an average of 95% of the students had achieved mastery. Only 5% of the students in the afternoon class had not attained mastery at the 80% level before moving on to the next unit.

Summary of Results

In summary, the t tests performed on the matched A.M. and P.M. experimental and control groups indicated that the groups were equivalent at the beginning of the school year, prior to the commencement of instruction. As a result, in computing effect size, no statistical procedures were required to account for initial differences in group means.

Both the morning and afternoon experimental group students performed higher than their control group counterparts on the teacher-made algebra test. However, the effect size of the afternoon group (.856) was much greater than the effect size for the morning group (.274) on the teacher-made test.

The effect size for the standardized test of algebra achievement differed between the groups. The afternoon experimental group achieved at a higher level than the afternoon control group (effect size .724). However, the morning control group performed at a slightly higher level (effect size -.181) than the afternoon experimental group.

On the whole, the students in the experimental groups were consistently more homogeneous on both posttests than the students in the control groups.

Conclusions

The conclusions to be drawn from this study should be held significant in light of the current debate over the effectiveness of group-based mastery learning as an instructional technique to enhance student learning. Moreover, the polarity of the coverage versus mastery issue holds implications for states and school districts who are involved in competency (specific mastery of skills) and normative-referenced achievement (broad coverage of content) testing as indicators of accountability. The conclusions to be drawn from this study can best be discussed in terms of the claims made by the proponents of mastery learning theory.

The Strong Claim for Mastery Learning - Standardized Test Claim

The strong claim for mastery learning contends that mastery learning students will achieve higher on standardized measures of achievement than those taught by traditional instruction when instructional time and learning objectives are held constant for both groups. In providing evidence for this claim, Slavin (1987) identified 7

studies that met the inclusion criteria and reported a median effect size of essentially zero (.04). Effect sizes included in the best-evidence synthesis ranged from essentially zero to .25. It is important to understand, though, that these effect sizes represent the mean effect for the total experiment. When several effects were included in any given experiment, the effect sizes were averaged. In this study they were not.

In analyzing the standardized test data from this study, contrasting effect sizes were found. An effect size of .724 was found in favor of the afternoon experimental group while an effect size of -.181 was found in favor of the morning control group.

If one were to look only at the afternoon groups, the strong claim for mastery learning could be clearly and powerfully substantiated. However, the data resulting from the morning groups apparently support the inverse; students in the control group performed higher than the students in the experimental group on the standardized measure of achievement.

A likely explanation for this phenomenon can be found in the results of the cycles of formative and summative testing for the experimental groups. An average of 84% of the students in the afternoon experimental class achieved mastery at the 80% level on the formative tests of the feedback-corrective cycles. However, an average of only 67% of the students in the morning experimental class achieved mastery at the 80% level on the formative tests. As a result, more time was required in the morning class to remediate deficiencies in prescribed skills. Since the two experimental groups moved through the curriculum together, more time was available for enrichment activities in the

afternoon class. Students were given additional drill and practice exercises as well as related work which broadened and extended their mastery of the content. It is most likely that this additional time for supplemental instruction was a contributing factor in enabling the afternoon experimental class to achieve higher on the standardized test than the control group, resulting in the .724 effect size.

It also must be noted that, while the performance of the morning and afternoon experimental groups was similar (means = 18.429 and 18.760), the achievement of the morning and afternoon control groups differed substantially. The mean of the morning control group was 18.947, while the mean of the afternoon control group was only 16.619.

The Curricular Focus Claim - Teacher-Made Test Claim

The curricular focus claim holds that mastery learning is an effective means of increasing student achievement of specific skills or concepts held to be the critical objectives of the course of study. The claim asserts that mastery learning students will perform higher on a teacher-made criterion-referenced measure of achievement than students taught under traditional learning conditions. Slavin identified eight studies which met the inclusion criteria and reported a median effect size of .255. The scores ranged from .18 to .90.

The effect sizes reported in this study were .856 in favor of the afternoon experimental group and .274 in favor of the morning experimental group. The range of effect sizes achieved in the study closely resemble the range of effect sizes reported by Slavin.

The curricular focus claim for mastery learning is substantiated by this study.

The Extra Time Claim

The validity of the extra time claim was not addressed by this study. The extra time claim contends that mastery learning is an effective use of additional time and instructional resources to bring almost all students to an acceptable level of achievement. In this study, no additional time beyond the regularly scheduled class period was made available for remedial or corrective instruction. Since instructional time was held constant for all groups in this study, no comparisons can be made to the best-evidence synthesis.

It is, however, the issue of extra time that is at the heart of the mastery learning controversy. In this experiment, time was held constant. In other words, no extra time beyond the regularly scheduled class period was given to students in the experimental groups to enable them to achieve mastery at the 80% level.

It is critical to remember that the learning for mastery theory predicts that most students can learn the material being taught when given sufficient time to do so. However, it can be argued that in any setting students who are given more time to learn will achieve at a higher level. The methodological decision to hold time constant in this experiment reflects the experimenter's desire to test mastery learning theory under Slavin's "best evidence" conditions and to draw conclusions in light of that methodological choice.

With that in mind, the results of this study can be compared to the findings of recent comprehensive reviews of the literature.

Comparison With Other Literature Reviews

The Block and Burns (1976) literature review reported a mean effect

size of .83 based upon 97 LFM comparisons of Type 1 research, i.e. research comparing the end-status achievement of students, variability, and types of learning. The 97 studies included both standardized and locally constructed posttest achievement measures. Kulik et al. (1986) report an average effect size of .52 on achievement as measured by final examinations (criterion tests) in 34 comparisons. Guskey and Gates (1986) reported a range of effect sizes from .02 to 1.70 based upon 34 studies in their meta-analysis. Only three of the 34 studies made use of standardized posttest measures; the remainder used teacher-made or experimenter-made posttests. In looking at effect sizes that are relevant to this study, Guskey and Gates report a mean effect size of .72 for high school students and an effect size of .81 for mathematics classes when effect sizes were sub-grouped from 25 comparisons.

None of the mean effect sizes reported in any of the four literature reviews approaches the 1.0 effect size reported by Walberg (1984).

Further Discussion

Even though effect sizes of .274 and .856 were found on the criterion-referenced measure, the raw score differences, .674 and 2.088 respectively, are minimal, especially to the eye of the classroom teacher. Similarly, with effect sizes of -.181 and .724 on the standardized posttest, the raw score differences, .518 and 2.141 respectively, are negligible for practical purposes. Looking at raw scores differences alone, the groups are very similar. However, the standard deviations of the group scores point out important differences in the relative homogeneity of the students themselves within the experimental and control groups.

In all instances, the experimental group students were more homogeneous than control group students on both of the posttest achievement measures. This homogeneity is a desired result of mastery learning. Mastery learning hopes to not only result in higher achievement for students, but also to result in more homogeneity within the group of learners. Thus, this study supports the premise that mastery learning is effective in not only increasing achievement levels, but also in making learners more alike in learning.

The Robin Hood Effect

One of the criticisms of mastery learning is that it fosters the dilemma known as the Robin Hood effect. This phenomenon occurs when instructional time is shifted from the faster learners to the slower students. Slavin (1986, p. 206) reports such findings in the best-evidence synthesis:

> ". . .given that overall achievement means are not greatly improved by group-based mastery learning, the reduction in standard deviations routinely seen in studies of these methods . . . are simply statistical indicators of a shift in achievement from high to low achievers."

The data from the morning experimental and control groups on the teacher-made algebra test tend to support the Robin Hood notion. While an equal number of students in both groups scored at the high end of the range, four times as many control group students scored at the low end of the range. It appears that mastery learning strategies were effective in raising the scores of the slower students in the morning experimental group while the scores of the faster-moving students, by contrast, remained constant.

The data from the afternoon groups on the teacher-made test do not, however, indicate an occurrence of the Robin Hood effect. While mastery learning did appear to raise the scores of the lower achievers (there were three times as many lows in the control group), there were concurrently three times as many high achievers in the experimental group.

A possible explanation for this occurrence, again, is that more time was available in the afternoon experimental class for enrichment. The combination of formative and summative testing to achieve the 80% mastery level while at the same time allocating sufficient time for enrichment may explain how the scores of lower achievers were raised without an occurrence of the Robin Hood effect taking place. However, by contrast, a strikingly different conclusion can be drawn from the standardized test data.

It is most enlightening to look at the variability of the students on the standardized test. In the morning groups, experimental and control students were similar at the higher end of the range. However, there were three times as many experimental group students at the lower end of the distribution. In this instance, mastery learning did not appear to be effective in raising the scores of the slower learning students on the standardized test. Furthermore, the control group students at the high end of the range performed higher than their experimental group counterparts.

On the other hand, an inverse relationship was found in the afternoon groups on the standardized test. The higher achievers in the experimental group did perform better than their control group counterparts (three times as great) while the scores of the experimental group

lower achievers were raised. Again, it should be noted that the mean of the afternoon control group was substantially lower than the mean of the morning control group, while the experimental group means were similar and that more time was available in the afternoon experimental group for enrichment.

Recommendations

The accountability movement in education has seen many states (e.g. Ohio) move in the direction of required courses of study which focus teachers on specific objectives, legislated criterion-referenced competency testing of students, mandated instructional intervention for those students who do not achieve competency, and mandated normative-referenced testing at several grade levels. Group-based mastery learning (which embodies the concepts of focused objectives, competency testing, and intervention) can be recommended to these schools as an effective way to raise levels of achievement on criterion-referenced measures of achievement.

However, school districts who focus instruction on a narrowly-defined set of objectives leading specifically toward criterion-referenced competency and proficiency should closely monitor the performance of students on normative-referenced measures of achievement. This study would suggest that when instructional time is focused on a detailed list of objectives and when no additional time is allocated for enrichment, scores on standardized tests suffer.

School districts who require instructional intervention for students who do not achieve competency on stated objectives must keep in mind that this intervention requires teachers to "rob" instructional

time from the "rich" in order to help the "poor." Instructional strategies need to be put into place that allow higher achievers to enrich and broaden learning concurrently with the remediation of lower achievers. Several strategies could be of use here, such as in-class groupings, cooperative learning, peer tutoring, and other techniques designed to remediate the learning deficiencies of lower achievers without appropriating substantial time from the higher achievers in the class.

BIBLIOGRAPHY

Block, J. H. and Anderson, L. W. (1975). Mastery learning in classroom instruction. New York: Macmillan.

Block, J. H. and Burns, R. B. (1976). Mastery learning. In L. S. Shulman (Ed.), Review of Research in Education (Vol. 4, pp.3-49). Itasca, Illinois: F. E. Peacock.

Bloom, B. S. (1968). Learning for mastery. Evaluation comment (Vol. 1, No. 2). Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation of Instructional Programs.

Bloom, B. S. (1974). An introduction to mastery learning theory. In J. H. Block (Ed.), Schools, society and mastery learning pp. 3-14. New York: Holt, Rinehart & Winston.

Bloom, B. S. (1980). The new direction in educational research: Alterable variables. Phi Delta Kappan, 62, 382-385.

Bloom, B. S. (1984a). The two-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13 (6), 4-16.

Bloom, B. S. (1984b). The search for methods of group instruction as effective as one-to-one tutoring. Educational Leadership, 41 (8), 4-17.

Burns, Robert D. (1979). Mastery learning: Does it work? Educational Leadership, 37 (2), 110-113.

Carroll, J. B. (1963). A model of school learning. Teachers College Record, 64, 723-733.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, California: Sage Publications.

Guskey, T. R., & Gates, S. L. (1985). A synthesis of research on group-based mastery learning programs. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. Educational Leadership, 43 (8), 73-80.

Kuhn, C. J. (1985). A quasi-experimental study of mastery learning strategies in the teaching of intermediate French in a suburban high school. Unpublished doctoral dissertation, Miami University.

Kulik, C. L., Kulik, J. A., & Bangert-Drowns, R. L. (1986). Effects of testing for mastery on student learning. Paper presented at the annual meeting of the American Educational Research Association, San Fransico.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 15 (8), 5-11.

Slavin, R. E. (1987). Mastery learning reconsidered. Review of Educational Research, 57 (2), 175-213.

Walberg, H. J. (1984). Improving the productivity of America's schools. Educational Leadership, 41, 19-27.